International Journal of Knowledge and Systems Science
Volume 15 • Issue 1

A Review on Speech Recognition for Under-Resourced Languages:
A Case Study of Vietnamese

Trung-Nghia Phung, Thai Nguyen University of Information and Communication Technology, Vietnam*
Duc-Binh Nguyen, Thai Nguyen University of Information and Communication Technology, Vietnam
Ngoc-Phuong Pham, Thai Nguyen University, Vietnam

ABSTRACT

Fundamental speech recognition technologies for high-resourced languages are currently successful to build high-quality applications with the use of deep learning models. However, the problem of "borrowing" these speech recognition technologies for under-resourced languages like Vietnamese still has challenges. This study reviews fundamental studies on speech recognition in general as well as speech recognition in Vietnamese, an under-resourced language in particular. Then, it specifies the urgent issues that need current research attention to build Vietnamese speech recognition applications in practice, especially the need to build an open large sentence-labeled speech corpus and open platform for related research, which mostly benefits small individuals/organizations who do not have enough resources.

KEYWORDS

Deep Learning, DNN-HMM, End to End, Machine Learning, Speech Corpus, Speech Recognition, Under-Resourced Languages, Vietnamese Speech Recognition

INTRODUCTION

Speech recognition is a type of problem in the field of pattern recognition, so there are difficulties similar to other recognition problems. There are also other problems with the random change nature of speech signals. So far, there are five classic and major problems affecting the accuracy and performance of a speech recognition system (Tebelskis, 1995; Duc, 2003; Jurafsky, 2008; Lei, 2006; Yu & Deng, 2016) which include the speaker-dependent problem, co-articulation problem, vocabulary (dictionary) size problem, noise problem, language-dependent problem.

Each speaker has a different structure of the sound articulators, so the characteristics of the voice that are emitted are greatly influenced by the speaker. Even when a speaker pronounces the same sentence, the voice speaking out can be different due to the amount of air escaping from the

DOI: 10.4018/IJKSS.332869 *Corresponding Author

lungs, emotional status, health status, etc. In terms of speaker-dependent characteristics, speech recognition can be divided into two types. The first type is speaker-dependent speech recognition, which is specifically developed to work with the voices of only one or a few people. The second type is speaker-independent recognition; that is a recognition system built to recognize the voice of any person. Normally, the speech recognition error rate of a speaker-independent system is usually higher than the speaker-dependent speech recognition system.

In a continuous pronunciation sequence, each sound is often greatly influenced by the sounds preceding and following it. Therefore, words with discrete pronunciation in recognition will have higher accuracy than words in a continuous pronunciation sequence. Since the recognition quality for a continuous pronunciation sequence is also dependent on detecting boundaries and silences between two words. When the speaker pronounces at a high speed, the silences and boundary between words will be narrowed, leading to each word segment being confused or overlapping, affecting the accuracy of word recognition.

The dictionary size is the number of all the different words that a particular recognition system is capable of recognizing. The larger the size of the dictionary is, the higher the complexity of the recognition system will be. The error rate of the recognition system is always proportional to the size of the dictionary. The speech recognition systems applied in practice today mostly require a large dictionary that covers all phonetic units to be able to recognize any sentence or word. These recognition systems require the training speech corpus to be large enough and cover all phonetic units such as phonemes in different contexts. While high-resourced languages such as English already had many labeled and widely used large speech corpora with hundreds of Terabytes, the large speech corpus problem for under-resourced languages including Vietnamese is still an unsolved problem.

In practice, the speech signal is often affected by noise from the outside environment such as traffic, animals' sounds, or the voices of one or more other people speaking at the same time. For humans, distinguishing and focusing on a person who is speaking to understand and distinguish semantics is quite simple, but for computers such cases cause special difficulties for identification because microphones pick up every type of audio signal in the frequency band in which it operates. Currently, even when applying optimal preprocessing methods on the received signal, and at the same time filtering out the signal of the speaker that wants to be identified, the recognition quality for these cases is still very low.

Each language has its own set of characters and phonemes. Researching and finding a set of standard phonemes for a language will improve recognition accuracy. For each language, the grammatical problem of pronunciation also greatly affects the quality of recognition. Pronunciations that follow a clear and complete syntactic structure are more accurately recognized than free pronunciations; that is words in pronunciation without specific grammatical constraints.

In this study, the authors review the results of fundamental research on speech recognition in general to handle the above 5 existing problems of speech recognition in Section 2; review research results on Vietnamese speech recognition - an under-resourced language in Section 3; and make recommendations for further studies on Vietnamese speech recognition in Section 4.

FUNDAMENTAL RESEARCH RESULTS ON SPEECH RECOGNITION

Currently, there have been several scientific publications on many different aspects to contribute to improving the quality of speech recognition. There are many ways of classifying studies on speech recognition. This paper reviews current studies based on four main components of a recognition system (Yu & Deng, 2016) including: Feature extraction; Acoustic model; Language model; The decoder, as well as the review of the latest research trends on the outstanding role of statistical learning methods and large corpus in speech recognition that aims to solve these five existing problems of speech recognition mentioned in section Introduction.

Extracting Acoustic Features

In the early stages of research on speech recognition, there have been many research results on speech acoustic features to improve the quality of speech recognition. The proposed speech features also aim to solve the problems of speaker dependent problem, noise problem and language dependent problem.

Two features that are commonly used in large vocabulary continuous speech recognition systems are Mel-frequency cepstral coefficients (MFCC) and Perceptual Linear Prediction (PLP) (Muda, Begam, & Elamvazuthi, 2010; Florian et al., 2005). These are acoustic features suitably proven for speech recognition systems and robust to noise. Studies of enhancement of the quality of features typically are improvements based on these two basic features. Advanced techniques are generally to find a transformation model or classification model to transfer these two features to a new spatial domain that increases the difference between two samples in two different classes to increase the quality of the recognition system. Some popular techniques include the following:

- **Linear Discriminant Analysis (LDA) (Haeb-Umbach, 1992; Sakai, 2007):** This technique seeks to find a matrix that transforms input feature to an output feature so as to increase the linear relationship between samples in the same class. LDA is commonly applied as a preprocessing step to improve quality and reduce dimensionality for input features such as MFCC or PLP.
- **Maximum Likelihood Linear Transform (MLLT) (Psutka, 2007):** This method is often used in conjunction with LDA: MLLT also finds an input feature transformation matrix into a new spatial domain such that samples in the same class will be better modeled by Gaussian models. The Maximum Likelihood is the parameter to separate the classes in the process of finding the transformation matrix.
- **Speaker adaptation:** This technique is generally to find a separate transformation model for each speaker. Then the feature vector corresponding to each speaker will be transformed into a new space through that speaker's transformation model to filter and carry more information about that speaker. In practice, this technique significantly improves the recognition quality for the system. However, its disadvantage is that it only works well with speakers who already have a transformation model, the recognition for a new speaker requires new data to retrain the system. In the study (Anastasakos, 1997), the authors proposed a training method to find the speaker-dependent transformation matrices for the input features before inputting the distributed probability distribution function of the hidden Markov model, although this method has been proposed for a long time (in 1997), thus far many systems still apply or use techniques based on this method. In the study (Martin, 2011), the authors proposed to use the i-vector speaker-describing feature vector to train the phonetic model, which increases the absolute quality of recognition by about 0.8%.
- **Using neural networks for feature extraction:** This is a new method and the research results show that it can improve the system quality. Normally, neural networks are applied in classification problems. Then, the value at the output layer of the network can indicate the class to which the input feature belongs, or can indicate the probability that the input feature may belong to the classes of the system. However, this new approach uses the value of the activation function of a hidden layer in the network as a direct input feature value for the hidden Markov model. The return of neural networks in recent years, especially deep learning network techniques and the robust development of parallel computing technology based on Graphical Processing Unit (GPU) has promoted these studies and achieved many results (Gehring, 2013; Kevin, 2014). In these studies, the authors used a hidden multi-layer neural network with parameters initialized by unsupervised training to calculate a feature called Bottleneck. This type of feature on average improves quality at a rate of about 10%.
- **Regarding tonal features:** From recent studies, it has been shown that most of the used types of features are acoustic features calculated based on the input signal spectrum to represent the characteristics of phonemes in a language. This feature is very effective for non-tonal languages

like English and German. For tonal languages, that is tones combined with phonemes also create the semantics of words, phonetic features do not fully express this tonal information. Pitch is caused by vibrations of the vocal cords during pronunciation, which usually persists during a syllable's pronunciation. The methods of calculating the tonal characteristic are usually based on the fundamental frequency F0 of the input speech signal. Tonal features are commonly used in speech synthesis, but are not commonly used in speech recognition. One of the reasons is that tonal features require some additional preprocessing before being used since tones do not exist in the voiceless region of a pronunciation. Some studies (Kevin, 2014) have shown that the integration of tonal features with phonetic features increases the recognition quality by about 2% absolute. This shows that the study of applying tonal features, especially to tonal languages like Vietnamese, is a necessary research direction to improve the quality of the recognition system.

Acoustic Model

Acoustic modeling is an important issue affecting the quality of speech recognition systems, especially for dealing with speaker-dependent problems, co-articulation problems, and language-dependent problems.

Up to now, two basic types of learning models are commonly used in speech recognition: 1 – Hidden Markov model combined with Gaussian model (HMM-GMM); 2 – Neural network model (NN), and more recently end-to-end model. Current studies are mainly carried out on these two types of models or hybridized both types in one, 1 and 2. The methods mainly focus on optimizing the parameter estimation process for the model based on a particular set of training samples. There are many improved techniques that have been proposed focusing on the main directions such as speaker adaptive training, parameter estimation to optimize the autocorrelation value between samples in the same class (Maximum Likelihood), parameter optimization based on feature space adaptive training, and multi-stream model, subspace model. Some of the commonly used methods include the following.

In the study (Anastasakos, 1997), the author proposed a method to train acoustic models whose parameters are optimally estimated according to the speaker (Speaker adaptive training-SAT). This method is based on the HMM-GMM model. A transformation matrix is found based on the input speaker data and information. Then, the input feature will be transformed into the new space through this matrix before being fed into the GMM model. Features in the new spatial domain have been reclassified based on maximizing the relationship between vectors belonging to a particular speaker.

Research (Daniel, 2010) proposes a method to train the acoustic model in the case of limited training data. For probabilistic models, data is an important factor in estimating phonetic model parameters during training, the lack of data can lead to the model receiving only initial random parameters or cannot describe all possible cases of the input pattern. In fact, for some newly studied languages, there are often very limited data, even for languages that have been studied for many years, there are special cases where there is little data such as the appearance of a new speaker for the system or the system has to work with a new context or a new environment. The proposed model in this study can solve this problem. The main idea of the method is that all Gaussian models of the identity units in the system will share another Gaussian model. This model is called a sub-Gaussian model (SGMM) where its parameters are determined through all the parameters of the models of the identity units in the system. The author's tests have shown that it improves recognition quality by 9.7% on average.

Research (Tokuda, 1999) proposes a new type of hidden Markov model which is capable of modeling a feature type containing both numbers and symbols. This model is named Multi-space Probability Distribution Hidden Markov Model (MSD-HMM). As soon as it was proposed, the author applied it to speech synthesis. The author uses this model to model a feature type with two separate streams in which one is a phonetic feature containing real numbers, the other stream contains information about pitch. The special thing is that the tonal feature can contain both real numbers

and symbols. This method was then applied mainly in the field of speech synthesis (Yu & Deng, 2016; Kunikoshi, 2011) and speaker recognition (Miyajima, 2001). Although this can be temporarily considered as a feasible solution for tonal languages because this model is capable of accurately modeling the discontinuity of tonal features, so far there has been very little research applying this model for speech recognition. MSD-HMM has only been applied to Chinese Mandarin (Qian, 2009; Chong, Wen, & Bo, 2011).

In the research (Ochiai, 2018), the authors have proposed a new method using deep neural network model as an acoustic model, but the hidden layer in the middle of this model is retrained for each speaker. Then, for each specific speaker, the speaker-dependent model will be the other layers of the network combined with the hidden middle layer which has been trained for this person. The results show that the new model increases by about 8.4% compared to the speaker-independent model.

Deep Belief Network - Deep Neural Networks (DBN-DNNs) were introduced as the core technology used to do acoustic modeling (Mohamed, Dahl, & Hinton, 2011). It replaced the 30-year-old standard in the industry of speech recognition systems: the GMM. From then, hybrid model between a hidden Markov model and deep neural network model (DNN-HMM) were popularly used by prominent speech research groups like Google, IBM, Microsoft Research, etc.

Research results show that most new studies focus on only a few common languages. Most of these languages are non-tonal languages, so the tonal feature is either ignored or used only as a component to enhance the recognition quality. The broken segments of the tonal feature are compensated by a random value through smoothing or cross-correlation algorithms. Only study (Tokuda, 1999) deals with modeling this broken characteristic. However, this model has not been studied extensively in speech recognition for other languages.

Language Model

Language model is also an important issue affecting the quality of speech recognition systems, especially for dealing with coarticulation problems and language-dependent problems.

Currently, the methods of building language models are often based on two main techniques, n-gram model and neural network. The n-gram-based methods were developed at a very early stage and are still widely applied today due to the simplicity of the model. The main disadvantage of the model is that it cannot determine the probability of sequences of words or words that do not appear in the data. There have been many studies (Ney, 1995; Katz, 1987; Frederick, 1980; Good, 1953) to overcome this shortcoming, which is called the smoothing method. Some commonly used methods are: add-one smoothing; back-off smoothing method; interpolation smoothing; Kneser-Ney smoothing.

The second type of language model is based on the neural network model. This type of model is often better than the n-gram model because it takes advantage of the network's classification capabilities. However, it is common that to train this type of model requires more data and more memory. In recent years, this type of model has been developed by many researchers with a number of new improvements. Research (Schwenk, 2007) presents a method of using deep learning networks to make language models. In this study, the authors did many tests, showing that the model using deep learning neural networks gives better results than the n-gram model by about 1% on average.

Decoder

The research on the decoder is mainly aimed at speeding up the speech recognition systems, optimizing the real-time recognition system.

The decoders in speech recognition systems today are mainly based on the Viterbi search algorithm, which is essentially to find an optimal path from a graph where the vertices are the identity unit of the system and the path weights or transition probabilities between vertices are calculated from the language and phonetic models. Some recent studies have only introduced new techniques to increase search speed or reduce memory capacity A typical method that can be shown is the method using the Finite-State Transducer (FST) (Dixon, 2012). The idea of the method is to integrate and

represent language models, phonetic models, and dictionaries into a single state transformation model. Thus, when decoding from an input through the FST model, we can find the best path without having to recalculate on the language model, the phonetic model. This method minimizes the decoding time for the recognition system, which is very effective for online recognition systems.

Latest Research Trends on Speech Recognition Based on NN Models

In the last two decades, there has been widespread recognition that the size, quality of the speech corpus, and the appropriateness of machine learning methods are key factors of success in most stages of the speech processing (Vu, 2005; Nguyen & Phung, 2017; Phung, 2013; Do, 2016; Adams, 2016; Novitasari, 2018; Luong, 2006; Do, 2015a; Do, 2015b; Trmal et al., 2014; Do, 2016; Novitasari, 2018; Do, 2017; Deng, 2012; Wang, 2019; Do, Sakti, & Nakamura, 2018; Tran, 2020; Zehra et al., 2021; Bashir et al., 2021). All speaker-dependent problems, coarticulation problems, vocabulary (dictionary) size problems, noise problems, language-dependent problems can all be solved synchronously with machine learning methods using large data datasets. Therefore, the latest research on speech recognition in the world is focused on improving the quality of machine learning methods with large data corpora.

When DNN-HMM was introduced as the core technology used to acoustic modeling (Mohamed, Dahl, & Hinton, 2011), the standard in the industry of speech recognition systems was replaced from GMM to DNN. Speech recognition systems using DNN-HMM trained with large, labeled speech corpora give recognition results with high accuracy close to human ability and could have been developed for practical applications (Do, 2015a; Do, 2015b; Trmal et al., 2014; Do, 2016; Novitasari, 2018; Do, 2017; Deng, 2012). However, speech recognition systems using the DNN-HMM model still have unresolved shortcomings and limitations.

Firstly, the recognition systems using the DNN-HMM model require a long training time, leading to high costs for training and system development.

Secondly, the recognition systems using the DNN-HMM model are sensitive to the accuracy of data labels while the construction of large datasets labeled with high accuracy is very expensive and unsuitable for research groups, small and medium enterprises.

Third, although the accuracy of the current best DNN-HMM model recognition systems can be over 95% with clean speech, further improving the accuracy of the recognition system is very difficult because this accuracy has reached a threshold close to human ability that many studies have not been able to surpass even with the integration of improved solutions.

Fourth, the recognition systems that use the DNN-HMM model for both training and recognition have the weakness of low speed, approximately one-second delay for real-time recognition (for short recognition segments) and increased for long recorded file recognition systems. For real-time recognition systems, a delay of one second is acceptable in practical applications, but will be less efficient with large, recorded file recognition systems. For example, for transcribing recorded files of reporters, low recognition speed will cause difficulties and inconvenience for reporters when they need to work quickly. Studies to speed up training and recognition for DNN-HMM models have not yielded many results.

Because of the limitations of the DNN-HMM model, the latest research on speech recognition in the world is quickly turning to using end-to-end model instead of DNN-HMM model (Wang, 2019; Do, Sakti, & Nakamura, 2018; Tran, 2020). The end-to-end model can be divided into three different categories: connectionist temporal classification (CTC)-based, recurrent neural network (RNN)-transducer and attention-based. The latest results show that the training procedure using the end-to-end model is very simple, the training and recognition costs are low, and the dependence on the accuracy of the data labels is reduced.

End-to-end model can directly use sentence-labeled speech annotation corpus without using phoneme aligned data for training. Therefore, it is suitable to develop speech recognition systems for under-resourced language with end-to-end models since the cost to build a sentence-labeled speech

annotation corpus is lower than that of phoneme-labeled speech corpus, which is necessary to build HMM-GMM-based speech recognition systems.

While DNN-HMM speech recognition systems are difficult to implement in combination with GPUs to optimize parallelism due to the use of n-gram graph model (Deng, 2012), end-to-end learning model is very easy to implement on systems that use GPU instead of CPU to speed up recognition training (Wang, 2019).

Deep learning techniques have given rise to an alternative approach in the form of the end-to-end model. Unlike the HMM-based model, the end-to-end model utilizes a single model to directly map audio to characters or words. This approach eliminates the need for specialized engineering and instead relies on a learning process. Consequently, constructing and training the end-to-end model is simpler and requires no domain expertise. These benefits have rapidly made the end-to-end model a popular research direction in large vocabulary continuous speech recognition.

Although the accuracy of the recognition systems using the end-to-end model of the recognition systems for English was equivalent to that of the recognition systems using the DNN-HMM model, the speech recognition systems for under-resourced languages using the end-to-end model (like Vietnamese in (Tran, 2020)) currently have lower accuracy than systems using the DNN-HMM model. Therefore, researchers around the world are continuing to improve and develop end-to-end methods for under-resourced languages to gradually increase the accuracy of the recognition system close to the accuracy of systems using the DNN-HMM model.

Basically, the latest research results on speech recognition to date show that the fundamental technology for speech recognition is matured for the development of high-quality speech recognition applications in practice for high-resourced languages that have large, labeled speech corpora (like English which there are many large open speech corpora with over 1.5 TB of labeled audio that are widely and effectively used, such as TIMIT data corpus with 630 speakers, eight English - American dialects (Bashir et al., 2021).

Meanwhile, the problem of large, open speech corpora for under-resourced languages is still an unsolved problem (John et al., 1993). People talk a lot about the digital revolution that can connect the world, but currently only about 5% of the world's 6000 languages are connected and have high resources (Scannell, 2007), most of the world's languages have not been adequately interested, and even languages with millions of speakers may lack the resources needed to conduct research and develop new technologies.

Recently, there have been many studies that point to a number of risks to under-resourced languages as well as opportunities for under-resourced languages to be able to "borrow" technologies from high-resourced languages and development tools (Wet et al., 2017). However, it is complicated to build a large phoneme-labeled speech corpus, which is necessary to build HMM-GMM-based speech recognition systems, especially for under-resourced languages. Therefore, it is intricate to borrow HMM-GMM technology from high-resourced languages to under-resourced languages. The hybrid model DNN-HMM can use two phases in training with sentence-labeled speech annotation corpora. The first phase uses HMM as a forced alignment and the second phase uses aligned data for training with DNN. However, the recognition systems using the DNN-HMM model are sensitive to the accuracy of alignment algorithms.

Fortunately, the latest speech recognition technology, end-to-end optimizes loss function based on both input and output sequence and does not require forced alignment for training acoustic models. Therefore, speech recognition systems using end-to-end just require sentence-labeled speech annotation corpora, which can be built faster and easier than phoneme-labeled speech corpus. Consequence, the opportunities for under-resourced languages to be able to "borrow" technologies from high-resourced languages are practical and feasible with the latest speech recognition technology end-to-end.

Thus, the core problem to build high-quality speech recognition systems for under-resourced languages is to build a large sentence-labeled speech annotation corpus with several speakers, enough dialects and open to the research community. In addition to the main technical requirements to ensure

the size of these corpora, the accuracy of the data labels, the phonetic coverage is also an important criterion to ensure.

SPEECH RECOGNITION FOR VIETNAMESE: AN UNDER-RESOURCED LANGUAGE

See Table 1.

Table 1. Summary of popular corpus for Vietnamese speech recognition

Corpus	Sub dataset	Duration	Number Speaker	Style	Utterance	Unique syllable	Open/ Close
VinBigdata-VLSP2020-100h	vlsp2020_train_set_01	80h	NA	Reading and spontaneous	112514	5836	Open
	vlsp2020-ASR-T1-test		NA		7079	3353	
	vlsp2020_train_set_02	20h	NA		56427	7421	
	vlsp2020-ASR-T2-test		NA		18843	1358	
VIVOS	Training	15h	46	Reading	11660	4617	Open
	Test		19		760	1692	
FPT Open Speech Dataset (FOSD)		30h	NA	Reading	25921	7327	Open
VNSpeechCorpus	Common part	50h	50	Reading	37 paragraphs	1257	Close
	Private part	50h			2000 paragraphs	NA	

Vietnamese Speech Corpus

Most research groups on Vietnamese speech recognition still use their own closed corpora for experimentation. So far, there has not been a large and open Vietnamese speech corpus that has been officially published but some closed Vietnamese speech corpus were collected recently (Do et al., 2014). Table 1, lists some popular Vietnamese corpus for Vietnamese speech recognition task, such as the following.

In 2004, International Research Center MICA released the a speech corpus named VNSpeechCorpus consisted of about 100 hours of data collected from office and studio by 50 speakers (25 are females and 25 are males with 1257 unique syllables). This dataset has a diverse number of speakers of all 3 accents (North, Central, and South) and is well designed both in terms of dictionaries and distribution of acoustic units (Nga, Li, Li, & Wang, 2021). Each speaker is recorded 1 set of 45 minutes of common part and 15 minutes of private part, the dataset has a limited number of words However, this dataset is limited to internal studies, detail of unique syllables not public and does not open the corpus for the community.

In 2016, a speech corpus named AIlab VIVOS was collected with over 15 hours from about 50 Vietnamese native speakers with 12420 utterances (in 15 hours training set have 46 speakers (22 males and 24 females) with 4617 unique syllables and in 45 minutes testing set have 19 speakers (12 males and 7 females) and 1692 unique syllables (Le et al., 2004). However, this dataset is small in size and the speakers are all Southern.

In 2018, FPT Corporation released the FPT Open Speech Dataset (FOSD) open dataset. This dataset consists of 25,921 utterances manually compiled from 3 sub-datasets (approximately 30 hours in total, include 7327 unique syllables, speakers of all 3 accents North, Central, and South) (Tran,

2020). The dataset is useful for several speech-related research topics, including but not limited to text-to-speech... However, this dataset is small in size and the speakers in the reading style.

In 2020, VinBigdata released the VinBigdata-VLSP2020-100h open dataset, explicitly designed for the VLSP-2020 international workshop. It includes two speech styles: reading speech (around 20 hours with 56427 utterances and 7421 unique syllables) and spontaneous speech (about 80 hours with 112514 utterances and 5836 unique syllables) (Hieu & Quan, 2016). This is the largest published open dataset, carefully designed with a full set of training and test sets with a variety of speakers and topics such as news, stories, and Wikipedia.

Research on Speech Recognition for Numerals and Discrete Words

Research direction on speech recognition for Vietnamese numerals and discrete words was popular in the two past decades. In particular, the research studies were only performed on discretely pronounced speech, i.e., the delay between two consecutive words is large. The vocabulary number is only 10 in the case of numeral recognition, or less than 200.

Research (Duc, 2003) has proposed several methods of labeling for Vietnamese speech data with continuous pronunciation. At the same time, it was proposed to use a hybrid model between neural network and hidden Markov model for recognizing 10 Vietnamese numerals on phone recording data with recognition quality reaching 97.46% of word level. In this study, the author used continuous speech for testing. However, the author only focused on solving the problem of automatic data labeling and using the hybrid model to model the set of units in the numeral recognition problem. Research results show that using hybrid neural networks with HMM model gave better results than traditional HMM model.

In the study (VinBigdata, 2023), the authors researched and analyzed in detail the features and characteristics of the Vietnamese speech, such as phonemic and acoustical features, and tones. The research also presented the methods of feature extraction and analysis of the influence of noise. The two types of models used and compared in the research were neural networks and hidden Markov. The scope of the research only applied to discrete speech with 193 syllables. The pronunciation sentences have limited content for the problem of controlling some functions of electronic and computer equipment.

Research on Tone Recognition

These research studies only focus on recognizing tones in each discrete word, that is, the output of the recognition system is one of six Vietnamese tones. In spite of being applied on continuous speech, the research studies on this issue still use traditional models such as HMM or NN and tonal features are added with "artificial" values in the voiceless regions where it doesn't exist.

The research (Nguyen et al., 2018) presented the use of fundamental frequency F0 to characterize Vietnamese tones, then modeled by hidden Markov model to identify tones. A typical type of vector for tones based on the sum and difference of F0 and the energy value between two adjacent signal frames is proposed. From that result, the author has built a Vietnamese speech recognition system with integrated tone recognition for discrete words with an accuracy of about 94%.

The authors (Vu, 2008) researched Vietnamese tone recognition, but following the approach on continuous speech. In this research, the author proposed a typical type along with a method of normalizing it based on the fundamental frequency F0 and the energy value of the speech signal. The research results were tested on a dataset of medium-sized continuous speech. Tone recognition results reached 81.02%.

In the research (Vu, 2009), the authors proposed a method to recognize tones for Vietnamese using neural networks. The study also proposed a suitable, typical type and normalization method for the recognition model. The research results were tested on a dataset recorded from Vietnamese radio programs. The average 6-tone recognition quality is 83.83% dependent on speakers, about 2% higher than the system using the hidden Markov model.

Research on Vietnamese Large Vocabulary Continuous Speech Recognition

The world's major technology corporations such as Goggle, Microsoft have all provided speech recognition applications with large vocabulary continuous speech. In which, the most popular are multilingual speech translation systems that have been put into the application, and can be installed on users' smartphone systems. However, speech recognition products of the major technology corporations either do not have a Vietnamese version or have low quality as being applied to Vietnamese. For these products, we also do not master the technology and data platforms to be able to further develop and customize applications according to different specific requirements. The most successful Vietnamese speech recognition product so far among the products of major corporations is Speech-to-text by Google for Vietnamese speech recognition results with an average accuracy of 80 - 90%. It has advantages such as stability against noise and echo, multi-speaker recognition, multi-channel recognition, but also has some disadvantages such as low recognition accuracy with some Vietnamese dialects, no security due to a server located abroad, and no mastering of the technology. In particular, the biggest problem when using speech recognition products of major technology corporations is the problem of dialects that has not been effectively handled.

Because speech recognition products of major corporations still have limitations, there have been efforts by domestic units to research Vietnamese speech recognition over the years and some initial research results have been obtained.

That was one of the first research studies on Vietnamese large vocabulary continuous speech recognition (Dixon, 2012). In the research, the author presented the basic structure of Vietnamese and proposed to test some sets of phonemes with and without tones. Tests were performed on datasets recorded from Vietnamese radio stations using 2 typical types of MFCC and PLP, and the recognition model is hidden Markov. The recognition result reached 82.97%. Although the author did not use tone features, the results of applying the tone model are more optimal than of the non-tone phonemic model. From these results, it has been shown that tone is a factor contributing to the quality of Vietnamese recognition, similar to Mandarin and Cantonese.

The research (Vu, 2008) was one of the first research studies on Vietnamese large vocabulary continuous speech recognition with integrated tone model. The author presented a new approach to initialize the training of acoustic models for Vietnamese by inheriting acoustic models from equivalent phonemes of other languages. The author proposed the modeling of tones similar to the research (Dixon, 2012), i.e., adding tone symbols to the phonemic symbols in the phoneme set of the system. In the research, the author gave an approach to synthesize two types of acoustic and pitch into one to serve as input for the HMM model. The author also proposed a method to improve the language model by collecting more text data from Vietnamese websites. The test results achieved a word recognition error of 11%.

The research (Vu, 2009) exported a combination of phonemes between single phonemes and vowels to train an acoustic model for Vietnamese large vocabulary continuous speech recognition. The test results achieved an accuracy of 86.06% on a dataset size of 27 hours for training and 1 hour for testing. The research team focused on optimizing the phoneme set based on the pairing of basic phonetic units in Vietnamese syllables. Typical types and models were still traditional MFCC and HMM.

The research studies by MICA under Hanoi University of Science and Technology proposed a solution to inherit phoneme models of other languages such as English and French to train phoneme models for Vietnamese speech recognition (Nguyen & Vu, 2009), recommended libraries to build Vietnamese speech recognition systems based on YAST tool (Sethserey, 2010). In the research, the authors completely inherit the models of equivalent phonemes that have been trained in English and French languages to build a set of phoneme models for Vietnamese. This approach eliminated the difficulty of building a good enough training speech corpus. However, due to the use of similar acoustic models from English and French languages, which are non-tonal languages, the acoustic

model in this case has been unable to recognize tones. Tone recognition will depend entirely on the language model.

The recent study (Ferreira, 2012) presented an approach to build recognition systems for languages with limited training data. Vietnamese is one of the languages tested in this research. The author proposed the use of a common set of phonemes for tested languages. By inheriting the trained data or model for these phonemes to generate the model for a new language, the author also recommended a deep learning neural network which is an advanced technology being applied by many research studies today for feature extraction.

The author (Vu, 2014) studied on optimizing the phoneme set for Vietnamese language recognition. The author proposed and tested different sets of phonemes by combining the first, last, mono, and diphthongs with different combinations to find the set of phonemes with the best results on the dataset.

The latest research result of AILab, which belongs to the University of Sciences, Ho Chi Minh City National University, showed that the Vietnamese speech recognition system using a hybrid model of DNN-HMM with the Kaldi tool has a syllable error rate of 9.54% (corresponding to an accuracy of the recognition system of 80.46%) (Nga, Li, Li, & Wang, 2021).

FPT Technology Research Institute (FTRI) is also an institution that has many years of research on Vietnamese speech recognition. FTRI has updated the latest research trend on speech recognition to successfully develop a Vietnamese speech recognition system using both end-to-end and DNN-HMM models (Tran, 2020), in which the accuracy of the recognition system using the end-to-end model is 3.4% lower than that of the hybrid DNN-HMM model for clean speech and 7.6% for noisy speech.

Some Vietnamese Speech Recognition Products Have Been Applied in Practice

In the last 2 to 3 years, a number of big technology companies have also actively invested in research and product development in this field such as FSoft Software Company, Zalo Company, Viettel Group with research and applications on recognition and synthesis of Vietnamese for intelligent interaction, serving internal applications of these groups. Especially since mid-2018, with the formation of AI Research Institute and BigData Institute, it has attracted many human resources in the fields of computer science, artificial intelligence, and machine learning, including research and application into Vietnamese speech recognition and synthesis, promoting research in this area and for the application development of corporations such as VinFAST and VinSMART.

Startups such as VAIS, VBEE, InfoRe also actively participate in the market when providing speech recognition and synthesis products. Among them, the outstanding product automatic transcription of VAIS has been applied to a number of agencies and units (Nguyen, 2014). The common feature of most products of these enterprises is that they are just pure application of speech recognition and synthesis and have not really been closely followed and integrated into the operational processes of agencies and organizations. Therefore, the scope of application is not wide. VAIS's automatic transcription is an example. Although it has high stability, the quite high accuracy of the recognition system, in order to be really convenient for users, it is necessary to continue to develop features such as expanding the application on mobile platforms, embedded computers, or developing additional functions to support playback, storage and editing of recognized documents, support for searching, tracking, and reporting data transcription, system administration function.

CHALLENGES AND FUTURE DIRECTIONS IN VIETNAMESE SPEECH RECOGNITION AS AN UNDER-RESOURCED LANGUAGE

Traditional statistical learning methods used in speech recognition are based on GMM, like HMM-GMM. GMM-based methods are statistically inefficient for modeling data that lie on or near a nonlinear manifold in the data space. Therefore, GMM-based methods have been replaced by DNN-based methods in the industry of speech recognition systems. Several speech research groups, like

Google, IBM, and Microsoft have used DNN to do acoustic modeling in their speech recognition systems in the last decade. DNN has been considered as a significant invention for replacing the 30-year-old standard in the industry of SR systems – GMM. However, the speech community is seeing a fast trend of moving from DNN-based to end-to-end modeling for speech recognition. It can be said that, nowadays, end-to-end models are the state-of-the-art technology for speech recognition, and end-to-end methods are the core technology used to build a modern speech recognition system.

As an under-resourced language, Vietnamese speech recognition has been studied for two decades, and some important results have been proposed. However, these results are still limited.

The initial research on Vietnamese speech recognition started two decades ago only at the level of numeral recognition, recognition of discrete words, or modeling of Vietnamese tone recognition with discrete words. Over the past decade, there have also been a number of research studies on Vietnamese large vocabulary continuous speech recognition, and there have been some initial research results, but the research results on Vietnamese large vocabulary continuous speech recognition have not been close to the research results on popular high-resourced languages, leading to the limited application of research results on Vietnamese speech recognition in practice.

As mentioned earlier, research results in the last few years show the outstanding role of NN-based methods like end-to-end and large speech corpora in speech recognition. As a consequence, it is able to build high-quality speech recognition applications in practice for high-resourced languages that have large speech corpora.

And for the lexicon-free speech recognition system with only sentence-labeled annotation in the training dataset, the newest technology end-to-end can be "borrowed" for under-resourced languages such as Vietnamese speech recognition with advanced customization and development. However, the biggest challenge is still the need for a large and open sentence-labeled speech annotation corpus and open platforms for research and development. Currently, there is also no Vietnamese speech recognition platform that is open to the community to use. Therefore, a prerequisite for the Vietnamese speech recognition problem to be widely studied to contribute to the community as well as to deploy diverse and practical applications to contribute to society is the need to study the following problems: Building a large and open Vietnamese sentence-labeled speech annotation corpora; Building an open platform that inherits the newest technologies, such as DNN-HMM and end-to-end, that have been well applied to high-resourced languages, and develops and customizes it to suit Vietnamese; Building high-quality speech recognition software with features that are convenient for users using the large open Vietnamese speech corpora and the above open platform.

CONCLUSION

In this study, the authors review the results of fundamental research on speech recognition as well as research results on Vietnamese speech recognition - an under-resourced language, thereby making recommendations for further studies on Vietnamese speech recognition, especially the urgent need to build an open large Vietnamese sentence-labeled speech corpus and an open platform for related research.

ACKNOWLEDGMENT

This work is supported by the National Science Project under number KC-4.0-16/19-25.

REFERENCES

Adams, O. (2016). *Learning a Lexicon and Translation Model from Phoneme Lattices. EMNLP*, 2016.

Anastasakos, T. A. (1997). Speaker adaptive training: a maximum likelihood approach to speaker normalization. In Acoustics, Speech, and Signal Processing (ICASSP; pp. 1043 – 1046), Munich.

Bashir, M. F., Javed, A. R., Arshad, M. U., Gadekallu, T. R., Shahzad, W., & Beg, M. O. (2021). *Context aware emotion detection from low resource URDU language using deep neural network. Transactions on Asian and Low-Resource Language Information Processing, 2021*.

Chong, J. N., Wen, J. L., & Bo, X. (2011). *Prosody Dependent Mandarin Speech Recognition*. In *International Joint Conference on Neural Networks* (pp. 197-201), California, USA: IEEE.

Deng, L. (2012). *Scalable stacking and learning for building deep architectures. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Dixon, P. (2012). *Development of the SprinTra WFST Speech Decoder*. NICT Research Journal.

Do, Q. T. (2015a). *WFST-Based Structural Classification Integrating DNN Acoustic Features and RNN Language Features for Speech Recognition. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Do, Q. T. (2015b). *The NAIST SPEECH RECOGNITION System for the 2015 Multi-Genre Broadcast Challenge: On Combination of Deep Learning Systems Using a Rank-Score Function. SPEECH RECOGNITIONU*, USA.

Do, Q. T. (2016). *A Hybrid System for Continuous Word-level Emphasis Modeling Based on HMM State Clustering and Adaptive Training. Interspeech*, USA.

Do, Q. T. (2016). Transferring Emphasis in Speech Translation Using Hard-Attentional Neural Network Models. *Interspeech*.

Do, Q. T. (2017). Toward Expressive Speech Translation: A Unified Sequence-to-Sequence LSTMs Approach for Translating Words and Emphasis. *Interspeech*.

Do, V. H., Xiao, X., & Li, H. et al.. (2014). Cross-lingual phone mapping for large vocabulary speech recognition of under-resourced languages. *IEICE Transactions on Information and Systems*, *97*(2), 285–295.

Do, Q. T., Sakti, S., & Nakamura, S. (2018). Sequence-to-Sequence Models for Emphasis Speech Translation. IEEE Transactions on Audio, Speech, and Language Processing, 26(3), 545–557.

Doganer, A., & Calik, S. (2017). A New Approach Using Hidden Markov Model and Bayesian Method for Estimate of Word Types in Text Mining. [IJKSS]. *International Journal of Knowledge and Systems Science*, *8*(4), 17–29.

Duc, D. N. (2003). *HMM/ANN system for Vietnamese continuous digit recognition*. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 481-486). Springer.

Ferreira, E. (2012). *YAST: A Scalable SPEECH RECOGNITION Toolkit Especially Designed for Under-Resourced Languages. Asian Language Processing (IALP)*. IEEE.

Florian, H., Stemmer, G., Brugnara, F., ... (2005). Revising Perceptual Linear Prediction (PLP). In INTERSPEECH, Lisbon, Portugal.

Frederick, J. (1980). *Interpolated Estimation of Markov Source Parameters from Sparse Data. Pattern Recognition in*. North-Holland.

Gehring, J. (2013). Extracting deep bottleneck features using stacked auto-encoders. In Acoustics, Speech and Signal Processing (ICASSP; pp. 3377 – 3381), Vancouver.

Good, I. J. (1953). The population frequencies of species and the estimation of population. *Biometrika*, *40*(3/4), 237–264.

Haeb-Umbach, R. (1992). Linear discriminant analysis for improved large vocabulary continuous speech recognition. In Acoustics, Speech, and Signal Processing (ICASSP; pp. 13-16), California, USA.

Hayama, T., & Kunifuji, S. (2012). Information provision modules to support creation of slides with easily understandable presentation. [IJKSS]. *International Journal of Knowledge and Systems Science*, *3*(3), 26–41.

Hieu, L. T., & Quan, V. H. (2016). *A non-expert Kaldi recipe for Vietnamese Speech Recognition System. Proceedings WLSI-3 & OIAF4HLT-2*, Osaka, Japan.

John, S., Garofolo, et al. (1993). *DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1*. NASA STI/Recon technical report n 93.

Jurafsky, D. (2008). *Speech and Language Processing* (2nd ed.). Prentice Hall.

Katz, S. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. In Acoustics, Speech and Signal Processing (pp. 400 – 410), IEEE.

Kevin, K. (2014). *The 2014 KIT IWSLT Speech-to-Text Systems for English, German and Italian*. In *The International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe – USA.

Kevin, P. S. (2007). The Crúbadán Project: Corpus building for under-resourced languages. Building and Exploring Web Corpora*: Proceedings of the 3rd Web as Corpus Workshop,* 4.

Kunikoshi, A. (2011). F0 modeling and generation in voice conversion. In Acoustics, Speech and Signal Processing (ICASSP; pp. 4568 – 4571), Prague.

Le, V. B., Tran, D. D., Castelli, E., Besacier, L., & Serignat, J. F. (2004). Spoken and Written Language Resources for Vietnamese. *LREC*, *4*, 599–602.

Lei, X. (2006). *Modeling Lexical Tones for Mandarin Large Vocabulary Continuous Speech Recognition*. University of Washington.

Luong, C. M. (2006). *Design of Vietnamese Speech Corpus and Current Status. Proceeding of Chinese Spoken Language Processing, 6th International Symposium, ISCSLP 2006*, ISBN: 981-05-7009-0, pp 748-758.

Martin, K. A. (2011). iVector-Based Discriminative Adaptation for Automatic Speech Recognition. In Automatic Speech Recognition and Understanding (SPEECH RECOGNITIONU; pp. 152-157), Waikoloa: IEEE.

Miyajima, C. (2001). Speaker identification using Gaussian mixture models based on multi-space probability distribution. In Acoustics, Speech, and Signal Processing (ICASSP; pp. 433 – 436), Salt Lake City, UT.

Mohamed, A. R., Dahl, G. E., & Hinton, G. (2011). Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 14–22.

Muda, L., Begam, M., & Elamvazuthi, I. (2010). *Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. Journal of Computing, 2*(2), ISSN 2151-9617.

Ney, R. K. (1995). Improved backing-off for n-gram language modeling. In *Acoustics, Speech and Signal Processing* (pp. 181–184). IEEE.

Nga, C. H., Li, C. T., Li, Y. H., & Wang, J. C. (2021). *A Survey of Vietnamese Automatic Speech Recognition*. In *9th International Conference on Orange Technology (ICOT;* pp. 1-4), IEEE.

Nguyen, G. N., & Phung, T. N. (2017). Reducing over-smoothness in HMM-based speech synthesis using exemplar-based voice conversion. *EURASIP Journal on Audio, Speech, and Music Processing*, *2017*(1), 1–7.

Nguyen, Q. B., Mai, V. T., Le, Q. T., Dam, B. Q., & Do, V. H. (2018). *Development of a Vietnamese large vocabulary continuous speech recognition system under noisy conditions*. In *Proceedings of the Ninth International Symposium on Information and Communication Technology* (pp. 222-226).

Nguyen, T. C. (2014). *Automatic speech recognition of Vietnamese*. Technical University of Liberec.

Nguyen, T. C., & Vu, Q. (2009). *Advances in Acoustic Modeling for Vietnamese LVCSR. Asian Language Processing*. IEEE.

Novitasari, S. (2018). *Construction of English-French Multimodal Affective Conversational Corpus from Drama TV Series. LREC*, 2018.

Novitasari, S. (2018). *Construction of English-French Multimodal Affective Conversational Corpus from Drama TV Series*. LREC.

Ochiai, T. (2018). *Speaker adaptation for multichannel end-to-end speech recognition*. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP;* pp. 6707-6711), IEEE.

Phung, T. N. (2013). Improving naturalness of HMM-based TTS trained with limited data by temporal decomposition. *IEICE Transactions on Information and Systems*, *96*(11), 2417–2426.

Psutka, J. V. (2007). Benefit of Maximum Likelihood Linear Transform (MLLT) Used at Different Levels of Covariance Matrices Clustering in SPEECH RECOGNITION Systems. Text, Speech and Dialogue*, 10th International Conference (TSD)*, Czech Republic.

Qian, Y. (2009). A Multi-Space Distribution (MSD) and two-stream tone modeling approach to Mandarin speech recognition. *Speech Communication*, *51*, 1169–1179.

Quang-Hung, L. E., & Anh-Cuong, L. E. (2014). Syntactic pattern based Word Alignment for Statistical Machine Translation. [IJKSS]. *International Journal of Knowledge and Systems Science*, *5*(3), 36–45.

Sakai, M. C. (2007). Generalization of Linear Discriminant Analysis used in Segmental Unit Input HMM for Speech Recognition. Acoustics, Speech and Signal Processing (ICASSP; pp. IV-333 - IV-336), Honolulu.

Sakai, M. C. (2010). *Subspace Gaussian Mixture Models for Speech Recognition. Acoustics Speech and Signal Processing (ICASSP)*. IEEE.

Schwenk, H. (2007). Continuous space language models. *Computer Speech & Language*, *21*, 492–518.

Sethserey, S. (2010). *Unsupervised acoustic model adaptation for multi-origin non-native. INTERSPEECH*. IEEE.

Tanantong, T., & Parnkow, M. (2022). A Survey of Automatic Text Classification Based on Thai Social Media Data. [IJKSS]. *International Journal of Knowledge and Systems Science*, *13*(1), 1–25.

Tebelskis, J. (1995). *Speech Recognition using Neural Networks*. Carnegie Mellon University.

Tokuda, K. (1999). Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. In Acoustics, Speech, and Signal Processing (ICASSP; pp. 229-232), Phoenix, USA.

Trmal, J., Povey, D., & Khudanpur, S. et al.. (2014). *Improving deep neural network acoustic models using generalized maxout networks*. In *Proceedings of ICASSP* (pp. 215-219).

VAIS (2021). *An Automatic Transcription Solution*. Science and development (in Vietnamese). Retrieved September 23, 2021.

VinBigdata (2023). *VinBigdata shares 100-hour data for the community*. Vinbigdata.org. Retrieved April 01, 2023.

Vu, N. T. (2009). Vietnamese Large Vocabulary Continuous Speech Recognition. In *Automatic Speech Recognition & Understanding – SPEECH RECOGNITIONU* (pp. 333–338). IEEE.

Vu, N. T. (2014). *Automatic Speech Recognition for Low-resource Languages and Accents Using Multilingual and Crosslingual Information. Karlsruher - Germany: Karlsruher Instituts of Technologie – KIT*.

Vu, T. T. (2005). *Vietnamese Large Vocabulary Speech Recognition. Proceeding of the Eurospeech*, Lisbon, ISSN: 1018-4074.

Vu, T. T. (2008). Vietnamese tone recognition based on multi-layer perceptron network. Conference of Oriental Chapter of the International Coordinating Committee on Speech Database and Speech I/O System, Kyoto (pp. 253–256).

Wahlster, W. (Ed.). (2013). *Verbmobil: foundations of speech-to-speech translation*. Springer Science & Business Media.

Wang, D. (2019). An Overview of End-to-End Automatic Speech Recognition. *Symmetry*, *11*(8), 1018.

Wet, D., Febe, et al. (2017). Speech recognition for under-resourced languages: Data sharing in hidden Markov model systems. *South African Journal of Science*, *113*(1-2), 1–9.

Yu, D., & Deng, L. (2016). *Automatic speech recognition* (Vol. 1). Springer.

Zehra, W., Javed, A. R., Jalil, Z., Khan, H. U., & Gadekallu, T. R. (2021). Cross corpus multi-lingual speech emotion recognition using ensemble learning. *Complex & Intelligent Systems*, *7*(4), 1845–1854.

Trung-Nghia Phung received his Engineering degree in Electronics and Telecommunications from Hanoi University of Science and Technology (HUST) in 2002. He completed his Master of Science degree in Telecommunications from Vietnam National University –Hanoi (VNUH) in 2007 and his PhD degree in Information Science from Japan Advanced Institute of Science and Technology (JAIST) in 2013. He was Dean of Faculty of Electronics and Telecommunications, Head of Academic Affairs, and he has been Rector of Thai Nguyen University of Information and Communication Technology (ICTU). He has been a Vice President of Vietnam Club of Faculties-Institutes-Schools-Universities of ICT (FISU) and President of FISU Branch in the Northern Midlands, Mountains and Coastal Region of Vietnam. He was the recipient of the award for the excellent young researcher (Gloden Globe award) from Ministry of Science and Technology (MOST) of Vietnam in 2008. His main research interest lies in the field of interaction between signal processing and machine learning and he has published more than 70 research papers related to this field. He serves as a technical committee program member, organizing co-chair, program co-chair, track chair, section chair, editorial board member and reviewer of several conferences, journals and books. He is now an associate editor of Thai Nguyen University Journal of Science and Technology (ICT section).